# MATHEMATICS FOR DATA SCIENCE

November 29, 2024

Today: Last concentration inequalities
Application 1 : Johnson-Lindenstrauss

P.S. Looks like the exam will be on January 8 (Wednesday)

# ① More on subgaussian variables

$y$ is subgaussian $\Leftrightarrow$ $\begin{cases} \exists K>0, \ \mathbb{E}\left[\exp\left(\frac{y^2}{K}\right)\right] \leq 2 \\ \mathbb{P}(|y| \geq t) \leq 2\exp\left(-\frac{ct^2}{K}\right) \quad \forall t \geq 0 \end{cases}$

where $c>0$ is a universal constant that does not depend on $K$ nor $t$

$\|y\|_{\psi_2}$ : smallest $K$ that satisfies the definition

Ex) Rademacher, gaussian, ...

## Hoeffding's inequality for (sum of) subgaussian variables

Let $y_1, \dots, y_N$ $N$ independent, subgaussian random variables with zero mean ($\mathbb{E}[y_i]=0 \ \forall i$).

Then $\forall a \in \mathbb{R}^N$, $\forall t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i y_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\|a\|^2 \sum_{i=1}^N \|y_i\|_{\psi_2}^2}\right)$$

where $c>0$ is a universal constant that does not depend on $a$, $t$, or $\|y_i\|_{\psi_2}$

↳ Many other versions of the inequality

Ex) For variables with nonzero mean, apply the inequality to $y_i - \mathbb{E}[y_i]$

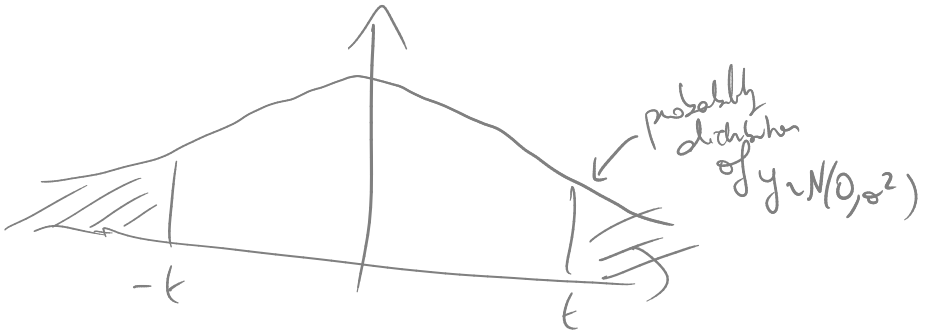
If $y_i$ is subgaussian, then $y_i + b$ is subgaussian $\forall b \in \mathbb{R}$

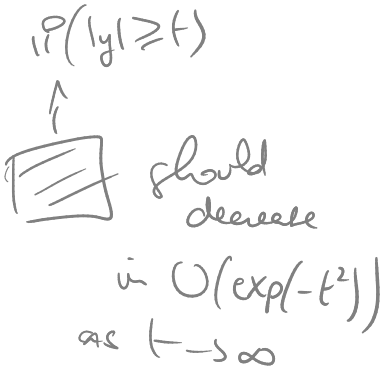
# (2) Subexponential random variables

↳ Many random distributions correspond to subgaussian variables but not all of them!

In particular, if $y$ is mean-zero gaussian $N(0,\sigma^2)$, then it is subgaussian, but $y^2$ is not.

$\Rightarrow$ Yet we expect $y^2$ to concentrate if $y$ does.




$$\mathbb{P}(|y| \geq t) = \mathbb{P}(y^2 \geq t^2)$$

**Def:** A random variable $y$ is called subexponential if $\exists K > 0$ such that

- $\mathbb{E}\left[\exp\left(\frac{|y|}{K}\right)\right] \leq 2$
- $\mathbb{P}(|y| \geq t) \leq 2\exp\left(-\frac{ct}{K}\right) \quad \forall t \geq 0,$

where $c > 0$ is a universal constant that does not depend on $K$ most.

The smallest $K$ that satisfies this property is denoted by $\|y\|_{\psi_1}$.

Subexponential $y$: $\forall t \geq 0 \quad \mathbb{P}(|y| \geq t) \leq 2\exp\left(-\frac{ct}{\|y\|_{\psi_1}}\right)$ for some $c > 0$

Subgaussian $y$ $\forall t \geq 0 \quad \mathbb{P}(|y| \geq t) \leq 2\exp\left(-\frac{ct^2}{\|y\|_{\psi_2}}\right)$ for some (other) $c > 0$

- For subexponential $y$, the probability $\mathbb{P}(|y| \geq t)$ decreases like $O(e^{-t})$ when $t \to \infty$, which is slower than $O(e^{-t^2})$ for subgaussian, but much faster than $O(\frac{1}{t})$ (Markov bound) or even $O(\frac{1}{t^2})$ (Chebyshev bound)

**Proposition:** $y$ is subgaussian $\Leftrightarrow$ $y^2$ is subexponential

Example. Chi-square distribution / Squared norm of a Gaussian vector, ...

**Def:** A random vector $y \in \mathbb{R}^n$ is subgaussian if

$v^T y \in \mathbb{R}$ is a subgaussian random variable for every $v \in \mathbb{R}^n$

$\Leftrightarrow (v^T y)^2 \in \mathbb{R}$ — subexponential —

Ex) . $y \sim N(0_{\mathbb{R}^n}, \sigma^2 I_n)$ is subgaussian

- $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ with $y_1, \dots, y_n$ independent subgaussian

is subgaussian

# Bernstein's inequality (Concentration for subexponential variables)

Let $y_1, \dots, y_N$ be independent, zero-mean, subexponential random variables.

Then, $\forall a \in \mathbb{R}^N$, $\forall t \geq 0$,

$$\mathbb{P}\left( \left| \sum_{i=1}^{N} a_i y_i \right| \geq t \right) \leq 2 \exp\left[ -c \min\left\{ \frac{t^2}{\|a\|^2 K_{max}^2}, \frac{t}{\|a\|_\infty K_{max}} \right\} \right]$$

where $K_{max} = \max_{1 \leq i \leq N} \|y_i\|_{\psi_1}$ and $c > 0$ is a universal constant that does not depend on $t, a$ or $K_{max}$

↳ Many other versions of the inequality

. Special case of interest: $a = \begin{bmatrix} 1/N \\ \vdots \\ 1/N \end{bmatrix}$

suggests that increasing $N$ decreases the bound (true for iid $y_i$'s)

$$\mathbb{P}\left( \left| \frac{1}{N} \sum_{i=1}^{N} y_i \right| \geq t \right) \leq 2 \exp\left( -c N \min\left( \frac{t^2}{K_{max}^2}, \frac{t}{K_{max}} \right) \right)$$

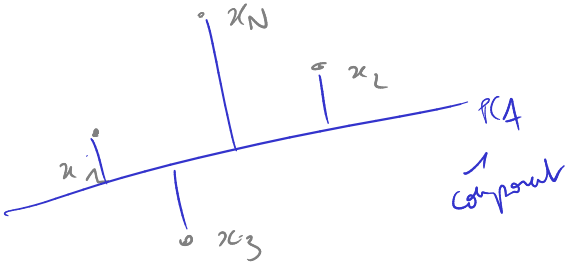
$$K_{max} = \max_{1 \leq i \leq N} \|y_i\|_{\psi_1}$$

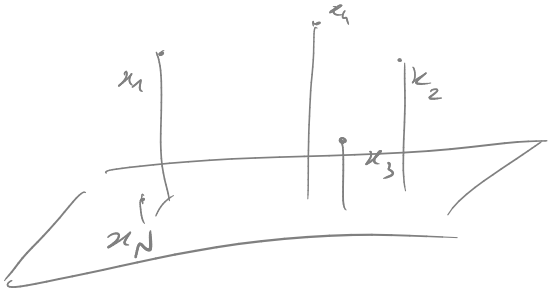
# (3) Application: Johnson-Lindenstrauss lemma

**Motivation:** Given $x_1, \dots, x_N$ $N$ vectors in $\mathbb{R}^n$ and $\varepsilon > 0$, we seek a low-dimensional subspace such that the projections $z_1, \dots, z_N$ of $x_1, \dots, x_N$ onto that subspace satisfy

$$\forall (i,j) \in \{1, \dots, N\}^2, \quad (1-\varepsilon)\|x_i - x_j\| \leq \|z_i - z_j\| \leq (1+\varepsilon)\|x_i - x_j\|$$

$\longrightarrow$ Techniques like PCA use information about the $x_i$s to define nested subspaces that best reflect the distances between the $x_i$s.




$\Rightarrow$ PCA requires calculations involving the $x_i$s

$\Rightarrow$ We want an alternative based on a priori definition of the subspace (without knowing the $x_i$s) and that works with high probability (concentration!)

**Theorem** (Johnson-Lindenstrauss lemma)

Let $\{x_1, \dots, x_N\}$ be $N$ points in $\mathbb{R}^m$ and $\varepsilon > 0$

Consider a matrix $P = \frac{1}{\sqrt{r}} \begin{bmatrix} y_1^T \\ \vdots \\ y_r^T \end{bmatrix} \in \mathbb{R}^{r \times m}$ $(r \leq m)$

defines a (random) subspace of dimension $\leq r$

with $y_1, \dots, y_r$ independent, zero mean, subgaussian random vectors such that $\mathbb{E}(y_i y_i^T) = I_m$ (ex: vector of iid Rademacher variables)

Then, $\exists\, C > 0$ universal constant (does not depend on $m, \varepsilon, N$) such that if $m \geq r \geq C \varepsilon^{-2} \log N$, (does not depend on $m$!) then

$$\mathbb{P}\left( \forall (i,j) \in \{1, \dots, N\}^2,\ (1-\varepsilon)\|x_i - x_j\| \leq \|Px_i - Px_j\| \leq (1+\varepsilon)\|x_i - x_j\| \right) \geq 0.99$$

($P$: projection onto the subspace; could be any $p \in [0,1)$ (affects $C$))

Desired property: Preserve distances up to $\varepsilon$

→ JL says that given $N$ points and an accuracy $\varepsilon$, there exists subspaces of dimension $O(\varepsilon^{-2} \log N)$ in which the projected points are at the same distances than the original points $\pm \varepsilon$ regardless of the ambient space $\mathbb{R}^m$ (with high probability)

→ The logarithmic dependency on $N$ comes from concentration, and is sometimes highlighted in stating the JL lemma:

" Given $N$ points $\{x_i\}$, $\exists$ subspace of dimension $O(\log N)$ such that the projections $\{z_i\}$ of $\{x_i\}$ satisfy

$$\forall (i,j), \quad 0.999\,\|x_i - x_j\| \leq \|z_i - z_j\| \leq 1.001\,\|x_i - x_j\|$$

with probability $0.99$ =

→ JL lemma is only interesting in high dimensions (so that $n >> O(\varepsilon^{-2}\log N)$ )

Proof sketch:

Showing the result is the same as showing

$$\mathbb{P}\left( (1-\varepsilon) \leq \|Pz\| \leq 1+\varepsilon \quad \forall z \in T \right) \geq 0.99$$

where $T = \left\{ \frac{x_i - x_j}{\|x_i - x_j\|}, \; x_i \neq x_j \right\}$

Key: $P = \frac{1}{\sqrt{n}} \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix}$ $\qquad$ $Pz = \frac{1}{\sqrt{n}} \begin{bmatrix} y_1^T z \\ \vdots \\ y_n^T z \end{bmatrix}$

$(y_i^T z)^2 - 1$ independent, zero-mean, subexponential

$y_i^T z$ subgaussian $\Leftrightarrow$ $(y_i^T z)^2$ subexponential

→ Applying Bernstein to $\{(y_i^T z)^2 - 1\}$ gives the result